



Enterprise Storage

White Paper

TECHNOLOGY OVERVIEW

Scale Computing's core offering is the Scale Storage Grid, or SSG. The Scale Storage Grid functions as an independently controlled mass data storage device, similar in function to a Storage Area Network (SAN) or Network Attached Storage (NAS) device.

The key difference between the SSG and the current state-of-the-art systems (SAN/NAS) is one of underlying architecture: The SSG utilizes both proprietary and licensed technologies that come primarily from the supercomputing (also called High Performance Computing, or HPC) sector. As such, the SSG consists of a number of small, specialized devices which, when combined with this supercomputing based technology, function as a single unit, delivering a storage network that provides a number of key, game changing advantages over previously existing technologies.

Detached Enterprise Storage (encompassing SAN, NAS, and similar systems) is a \$40 billion market, growing at 12% annually. The market is led by EMC, Hewlett Packard, Network Appliance, and DELL. Existing solutions rely heavily upon physically redundant components and RAID technology to address customer requirements.

CUSTOMER BENEFITS

Order of Magnitude Cost Savings

A traditional SAN system that delivers generally demanded functionality (high-availability, redundancy, connectivity) carries a price tag that is typically in the area of \$10,000 - \$20,000 per usable terabyte of storage capacity

The fundamentals of the SSG architecture drive down deployment costs. By utilizing a series of smaller devices rather than one (or few) large devices, redundancy and high availability are achieved without the need for the development of proprietary hardware. Instead, a combination of off-the-shelf or "commodity" hardware can be used in conjunction with software technology that manages the allocation of the data across all nodes, or "Storage Bricks" in the SSG. Each brick in the SSG contains a number of physical hard disk drives, and it is the underlying software technology, rather than proprietary hardware and controllers, which manages the distribution of data across both individual drives and across nodes in the grid.

The resulting cost savings is significant: The cost of deploying an SSG is \$2000-3000 per usable terabyte in a typical deployment, saving 50-80% over the costs of current SAN/NAS technology based solutions.

Significant Scalability Improvements over Current Technologies

Current SAN/NAS technology is very limited in its ability to scale up as storage requirements grow. The most common method of scaling involves the non-technology solution of forecasting future storage needs and grossly overbuying capacity today in order to have free capacity in the future. In some limited cases, a number of SAN devices can be interconnected, but in doing so some elements of high-availability are lost in that losing a single SAN component will take the entire array offline (see below). Further, when the device reaches capacity, the standard procedure is to purchase an entirely new device, migrate data, and repurpose/decommission the previously used device, often at great expense.

The SSG technology makes expanding capacity quite effortless and straightforward. Simply adding a new brick to the array will expand the capacity of the entire array, without the need to migrate data or take the old bricks offline. Hundreds or even thousands of bricks can be added on an as-needed basis, allowing customers to grow from small data stores of just a few terabytes, growing then up into the petabytes by simply adding additional bricks as they go.

Substantial Improvements in High Availability, Redundancy, and Recovery

The SSG architecture represents a paradigm shift in the ability to recover from hardware failure without the use of RAID technology. The underlying file system technology ensures that data is mirrored and redundant elsewhere in the grid, such that data recovery in the event of a failure is fast and painless. In the event of hardware failure, the controlling software instantly begins re-replicating the lost data so as to be prepared to recover from a potential subsequent failure. Redundant data copies are not kept on the same individual brick, thereby ensuring a smooth recovery even if an entire brick is taken offline suddenly and without warning. Likewise, the controlling software ensures that adequate disk space exists elsewhere in the grid for re-replication to take place. Data is distributed at the block level, providing for the ability to replicate even very large individual files (such as databases or video files) in this manner.

System uptime is maintained up to $N/2 - 1$ brick failures. For example, in a 100 brick grid, up to 49 devices can be generally be lost before the system is taken offline.

No Single Point of Failure

Other attempts to utilize a multiple device based architecture have been plagued with the problem of incorporating a single point of failure somewhere in the system. Typically, these systems rely on either a central “master” node for either configuration, data allocation, or both, which then issues instructions or sends data information to the “slave” nodes which are essentially passive receivers. This introduces the inherent problem of creating a single point of failure (the master node) at the point in where the system is under the most stress (all data requests, user permissions and/or configuration changes flowing through the master).

In the case of SSG, our proprietary grid control technology creates an architecture in which all nodes are active nodes and no single node functions as a master. This is best illustrated by the example of configuration management: an administrator can initiate changes to the entire grid by accessing a single administrative interface, and that interface is available from any of the individual nodes in the grid.

Likewise, any individual node on the grid can identify where other data physically exists on the grid without accessing any kind of master database (which would exist on a master node). It's in this way that re-replication of data is possible regardless of what specific node (or disk) failure triggered the need for re-replication.

Throughput and Performance Improvements

Connectivity into most SAN/NAS architectures consists of one or two inputs through which all data flows, regardless of the size of the system. As such, simultaneous connections to these devices result in throughput degradation. For example, if a traditional NAS has throughput capabilities of 50 MB/s, then that throughput will be shared across all simultaneous connections. Two connections will each receive half the speed, or 25 MB/s; five connections would each receive 1/5 the speed, or 10 MB/s; and so on.

The grid nature of the SSG architecture, combined with the elimination of a master node, results in the ability to deliver high throughput in a parallel access environment, because each node in the grid is yet another access point into the entire system. For example, if the SSG can deliver 50 MB/s and consists of 10 individual nodes, then 2 connections would each receive 50 MB/s; 5 connections would each receive 50 MB/s; and so on, until the point at which either (a) simultaneous connections exceeded the number of individual nodes in the grid, at which time throughput would be shared among them, or more likely (b) the total throughput exceeds the capacity of the backend network (gigabit ethernet, infiniband, fiberchannel) on which the grid is connected.

Furthermore, advancements in data storage protocols are making parallel access even more productive. For example, parallel CIFS (a standard data transfer protocol that is an improvement over standard CIFS) enables data read/writes to be performed across multiple devices simultaneously. If a server were using this protocol in conjunction with an SSG based storage array, massive performance improvements are attainable. For example, if a file was being written over standard CIFS on a traditional SAN, the throughput is limited by the maximum throughput of the SAN, and even more so by the number of parallel access connections to that SAN as previously described. However, using parallel CIFS with SSG, that single file can be written across all nodes simultaneously, such that, using the previous example, a 10 node grid could achieve 500 MB/s throughput (50 MB/s * 10 nodes) because each node is an autonomous entry point, effectively moving the throughput bottleneck from the storage system to the network backbone in most cases.

INTELLECTUAL PROPERTY

Masterless Grid/Node Architecture

The SSG architecture and the resulting Storage Bricks utilize a robust combination of licensed intellectual property, proprietary technology, and open source software. Among the proprietary components of the system is master-less grid/node architecture, specifically around the ability to change and control the entire cluster from any node – in essence, it is a grid of equals, where each node has the same abilities as any other node, and can be controlled, or initiate controls, as such.

Automation of HPC Filesystem Management

A number of key technology components are licensed, and most notable among these licenses is a license from IBM for a High Performance Computing (HPC) file system, a file system developed by IBM for use primarily in the supercomputing world. Installations of this HPC file system are typically unique to each customer, and involve substantial professional services work on the part of IBM to setup and then maintain these systems. A portion of our proprietary technology is the use of artificial intelligence based management systems which replace the need for customized professional service-centric installations. This uses an expert-system to handle the tasks which normally require human intervention in HPC file system management, including the complexities of node maintenance and disk provisioning. Utilizing this system, the time to complete a task such as drive provisioning is reduced from a day-long project requiring considerable expertise, into a simple point-and-click exercise taking 10 minutes or less.

AI Based Grid/Filesystem Management

The management system underlying the SSG architecture utilizes two key artificial intelligence based components. First, a proprietary expert system is utilized to handle the nuances of setting up and maintaining the grid architecture and the HPC file system. It is primarily due to this expert system that we are able to take a traditionally customized and complex network architecture (grid computing), and utilize it in such a way that the end customers (and administrators) are shielded from that underlying complexity. A primary roadblock to utilizing grid computing technologies in widespread commercial deployments is this very complexity, and our expert system is a part of our competitive differentiation at the technology level.

The second component of our AI-based system is our Predictive Hardware Failure (PHF) technology. PHF uses combination of machine learning techniques, including neural networks, to forecast when various hardware components are likely to fail, while there is still time to migrate data and smoothly shutdown those components before sudden unexpected failure occurs. This is done by monitoring low-level kernel and device drive information in real-time, and identifying the patterns that indicate impending failure. The integration of hardware and software is key to enabling this technology, and patent claims around this technology are being pursued.